基于 NGAS 的多点观测数据存储与同步方法

石□明[1,3]，邓辉[2,3]，戴伟[3,4]，卫守林[31]，王锋[1,2,3,4]

（1.昆明理工大学管理与经济学院，云南 昆明 650093；2.广州大学 天体物理中心/物理与电子工程学院，广东 广州 510006；3. 昆明理工大学云南省计算机技术应用重点实验室，云南 昆明 650500；4. 中国科学院云南天文台，云南 昆明 650011）

摘要：平方公里阵列（Square Kilometre Array，SKA）望远镜建成后将会具有超高的灵敏度、超快的巡天速度以及宽视场，进而产生超海量的观测数据。在 SKA 天文台与各国区域数据中心间的海量数据同步/传输是当前 SKA 建设中的一个难点。SKA 先导项目使用的下一代归档存储系统（Next Generation Archive System, NGAS）在应用测试中存在效率低下，性能不足等问题。本文提出了一种基于 ZeroMQ 的数据存储与同步方法，通过采用更加高效的异步消息机制实现同步传输数据，回避了 NGAS 原有的采用 HTTP 协议的局限。实验结果表明新方法在平均数据归档存储效率方面比 NGAS 原有方法快了将近 40
，能够基本满足 10GB 带宽的全速传输需要，取得了较好的使用效果。

关键词：NGAS；SKA；存储与同步；海量数据；ZeroMQ

中图分类号：P161;TP311.1 **文献标识码：**A **文章编号：**1672-7673(2017)xx-xxxx-xx

0 引言

由中国、澳大利亚、南非、英国等国家共同参与建设的平方公里阵列望远镜（Square Kilometer Array，SKA）将是最大的天文实验装置，具有前所未有的灵敏度、巡天速度和视场[1-3]。SKA 望远镜由分布在澳大利亚西部沙漠上的工作频率为 50-350MHz 的低频螺旋对数天线和南非及南部非洲 8 个国家上的工作频率为 350MHz-15GHz 的高频蝶形天线构成，其总接收面积达到 1 平方公里[4, 5]。SKA 将产生超海量的观测数据，在 SKA1 阶段，每秒产生高达数十 TB 的原始数据，需要长期保存的科学数据每年新增 50~300PB，SKA2 阶段，每年新增的科学数据将会达到 SKA1 阶段时的 100
[3, 6]。

为了应对站址国当前面临的预算和数据量处理的限制等相关问题，SKA 提出建设区域数据中心，实现数据的异地存储与归档，并通过各国科学中心的建设推动科学研究工作的方案。这一目标的达成显然要求海量观测数据能够高速地从观测地（南非和澳大利亚）的数据中心同步传输、存储到区域数据中心，从当前的技术水平来看，这一需求也具有非常大的挑战。

下一代归档存储系统（Next Generation Archive System，NGAS）最初在 SKA 先导默奇森宽场阵列（MWA）中开发完成，是当前射电天文领域最为常用的一套成熟观测结果归档软件，系统采用 Python 开发，功能非常丰富，具有高度的移植性。NGAS 设计之初是为了解决 ESO 在 20 世纪末期面临的每天新增的 55 GB 观测数据进行高效且低成本的数据归档、处理、检索及同步的问题[7]。已经实际应用在 MWA 与美国麻省理工学院（MIT）和新西兰的惠灵顿维多利亚大学（VUW）[8]间的数据同步，也被欧洲南方天文台（ESO）用来归档管理产生的海量观测数据及同步存储到不同的站点[9-12]；阿塔卡马大型毫米波/亚毫米波阵列（ALMA）使用 NGAS 将收集的观测数据同步到美国、日本和德国的区域数据中心[11, 12]。

然而，在面对 SKA 这一类具有更高时效性要求的数据同步与归档需求时，NGAS 仍然面临着一些问题。远程数据同步效率是其中的一个重要方面，这其中的根本原因在于 NGAS 在同步传输数据的过程中使用基于 HTTP 的方式来同步传输数据，由于 HTTP 协议封装效率较低，导致整个数据传输性能较差。SKA-1 的数据同步量相对较少，采用 HTTP 协议封装应可以满足要求。但随着 SKA-2 的建设，数据量呈指数级增长，这个时候研究新的封装方法，提高效率就成为一种必然。

由于当前 SKA 正在最终设计评估阶段，部分需求还没有最终确定。因此对于 SKA 数据归档工作和远程同步的工作均在预研与测试阶段，本文的工作正是在这个方面开展的基础工作。为提高远程数据同步性能，本文针对我国建设区域数据中心的需要，进一步研究了基于 ZeroMQ 的多点观测数据存储与同步方法。

1 NGAS 的底层实现分析

1基金项目：国家重点研发项目（2016YFE0100300，2018YFA0404603）资助，国家自然科学基金（No. U1531132，U1631129，U1831204，11403009，11463003，11773012）资助，塞尔网络下一代互联网技术创新项目（No. NGII20170204）资助.

作者简介：石聪明，男，博士，研究方向：信息管理信息系统，Email:shicongming@astrolab.cn

通讯作者：王　锋，男，教授，研究方向：天文技术与方法，Email:wangfeng@astrolab.cn

NGAS 的核心是一个多线程并发 HTTP 服务器，NGAS 通过关系型数据库(RDBMS)来管理归档文件的元数据、订阅者信息、磁盘信息等。NGAS[2]实现了 STATUS、ONLINE、OFFLINE、ARCHIVE、SUBSCRIBE、UNSUBSCRIBE 等 20 多个自定义命令，这些命令的主要功能是实现基本的数据归档与检索、服务器端数据压缩和过滤、自动镜像数据、磁盘跟踪、离线数据传输、数据一致性校验、数据订阅（数据存储与同步）等功能。

1.1NGAS 数据同步功能

NGAS 的数据同步传输功能是通过 NGAS 的数据订阅线程来调度订阅者对应的数据发送线程来实现将数据从数据发布者同步传输给数据订阅者。数据发送线程的流程图如图 1 所示。

NGAS 数据发布方将一个数据文件传输给数据订阅方都需要经历如下过程：用 HTTP 协议封装数据文件、将封装好的数据文件发送出去、等待接收数据订阅方响应的成功存储数据文件的消息、接着处理下一数据文件。NGAS 在同步传输数据过程中需要数据发布方等待接收数据订阅方反馈消息导致整个数据传输性能较差，同时由于 HTTP 协议封装效率较低，也会导致整个数据传输性能较差。

2 多点观测数据存储与同步方法的改进

针对 NGAS 中的数据同步传输功能是基于 HTTP 实现，考虑到当前技术发展的主流趋势，本文提出了一种基于零消息队列（ZeroMQ[13]）来改进 NGAS 中的多点观测数据存储与同步方法的方法。改进的方法使用 ZeroMQ 中的 PUB-SUB 套接字组合来实现高效快速的数据同步传输与存储。然而，PUB-SUB 套接字的组合却具有这些问题：1）订阅方崩溃导致订阅数据丢失；2）订阅者取回消息很慢导致发布方的发布队列溢出而造成的数据丢失；3）网络超载导致数据丢失；4）订阅方加入太迟错失了发布方已经发布的数据[14]。

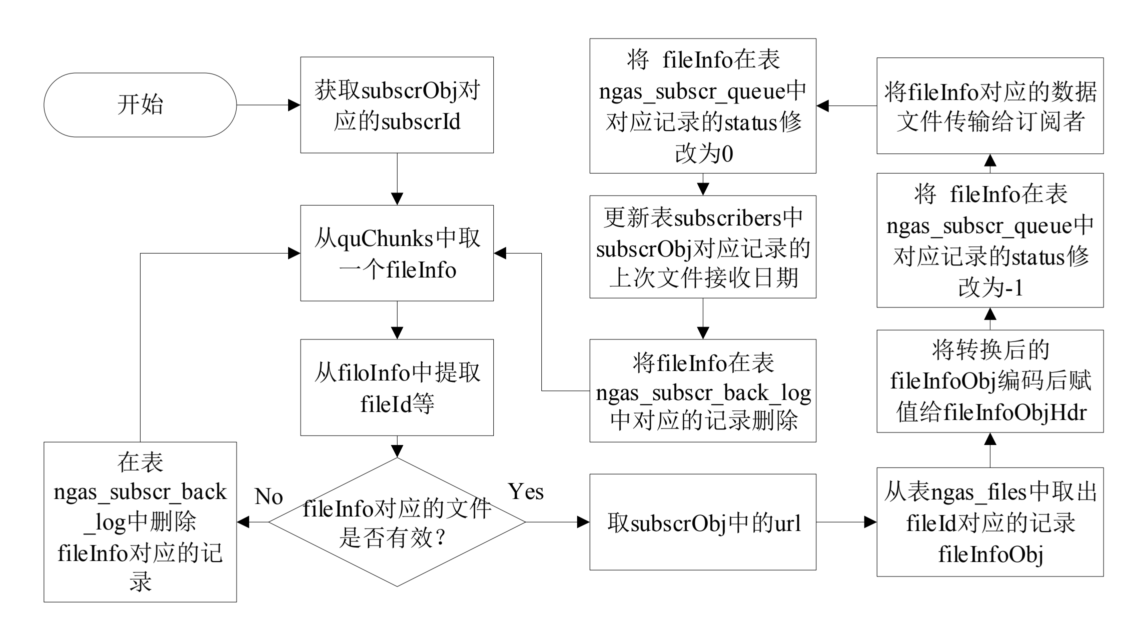


图 1NGAS 数据发送线程执行流程图

Fig.1A flowchart of data delivery

为了在改进方法实现的系统中解决 PUB-SUB 套接字组合带来的问题，我们加入了近实时感知端口连接状态的机制来规避发布方在没有订阅方连接的情况下发布数据，同时加入数据重发机制来克服因为网络超载、订阅方崩溃等导致的数据丢失造成的订阅方无法完全同步数据的问题。为了使基于改进方法实现的数据同步传输与存储子系统能够独立于 NGAS 运行，我们在该子系统中加入了使用 ZeroMQ 中的 DEALER 与 ROUTER 实现的订阅与退订功能模块。

基于改进方法实现的系统主要包括如下子系统模块：数据发布端服务器（Pub-Server）、数据订阅端服务器（Sub-Server）、订阅者服务器（Subscriber-Server）、订阅者客服端（Subscriber-Client）。Pub-Server 和 Sub-Server 负责数据发布方与订阅方之间的数据同步传输与存储，如图 2 所示；Subscriber-Server 和 Subscriber-Client 负责发布方与订阅方之间的消息订阅与退订，如图 3 所示。

2https://ngas.readthedocs.io/en/latest/index.html

2.1Pub-Server 与 Sub-Server 设计与实现

Pub-Server 主要负责启动数据发布端与数据同步传输与存储相关的守护线程，其执行流程图如图 4 所示。Pub-Server 主要包含如下功能模块：启动订阅者对应的发布数据守护线程、启动接收反馈消息的守护线程、启动处理反馈消息的守护线程、启动更新积压文件的守护线程、启动更新发布队列的守护线程、启动处理新增订阅者的守护线程、启动处理新增退订者的守护线程。Sub-Server 的功能与 Pub-Server 类似，只是处理对象不同。

Pub-Server 与 Sub-Server 之间的数据同步传输与存储中涉及到两种消息：Pub_msg 和 Sub_msg，如图 2 所示。Pub_msg 是数据发布方（Publisher）发布的消息，格式为 SI_SP:PI_PP:BFR:BFD；Sub_msg 是数据订阅方（Subscriber）发布的已成功接收与存储的反馈信息，格式为 SI_SP:PI_PP:BFR。SI 表示 Subscriber 的 IP；SP 表示 Subscriber 为某个 Publisher 申请的用于发布反馈消息的固定端口；PI 表示 Publisher 的 IP；PP 表示 Publisher 为某个 Subscriber 申请预留的用于发布数据的固定端口；BFR 由文件名、文件 ID、文件版本、文件类型组成；BFD 表示 BFR 对应的积压文件数据。

当 Pub-Server 上的近实时端口连接状态守护线程能检测到某个 Publisher 的数据发布端口被 Subscriber 连接上时，触发该 Publisher 对应的数据发布线程开始产生并发布 Pub_msg；否则触发停止数据发布线程。同时，数据重发机制会在某个 Publisher 发布某个数据文件超过一段时间仍未收到 Subscriber 发布的已成功接收和存储的反馈信息时，将让 Publisher 重新向 Subscriber 发布该数据文件。

2.2 Subscriber-Server 与 Subscriber-Client 设计与实现

Subscriber-Server 与 Subscriber-Client 之间的异步通信模式如图 3 所示。Subscriber-Server 负责接收处理订阅消息（Sub-Msg）和退订消息（Unsub-Msg），并将订阅成功消息（Sub-Msg-S）、订阅失败消息（Sub-Msg-F）或退订成功消息（Unsub-Msg-S）回复给对应的请求者，同时更新数据库中的相应订阅者记录；Subscriber-Client 负责向 Subscriber-Server 发送 Sub-Msg 和 Unsub-Msg，根据接收到的响应消息来更新数据库中的相应发布者记录。其中，Sub-Msg、Unsub-Msg、Sub-Msg-S、Sub-Msg-F、Unsub-Msg-S 这 5 种消息的格式分别为：S_SI_SP_Datetime、U_SI_SP、SS_SI_SP_PI_PP、SF_SI_SP、US_SI_SP。同时，我们在 Subscriber-Client 中加入了消息重发机制来确保 Subscriber-Client 能够成功订阅或者退订相应的数据发布者。

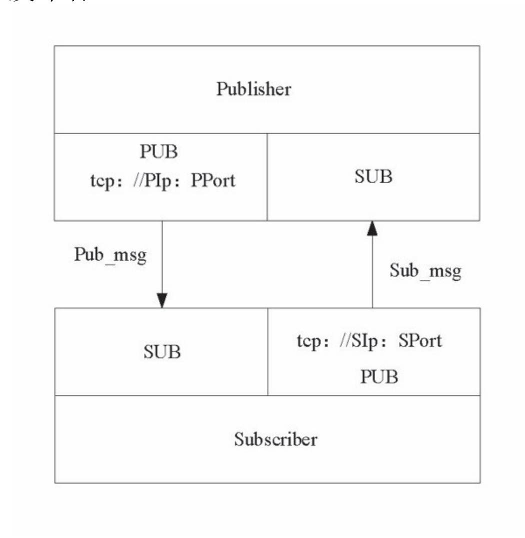


图 2 Pub-Server 与 Sub-Server 之间的通信模式

Fig.2Communication mode between Pub-Server and Sub-Server

Subscriber-Client
Client1 Retry Dealer
Client2 Retry Dealer
ClientM Retry Dealer
connect connect connect
bind
ROUTER Server Dealer
bind
connect connect connect
Dealer Worker1
Dealer Worker2
Dealer WorkerN
Subscriber-Server

图 3Subscriber-Server 与 Subscriber-Client 之间的通信模式

Fig. 3 Communication mode between Subscriber-Server and Subscriber-Client

3 实验

3.1 实验环境

本文测试性能所用的硬件环境为：1 台型号为 IW4200-10G 的思腾合力 GPU 服务器，该服务器具有 16 个双核 Intel®Xeon(R) CPU E5-2620 v4 @ 2.10GHz 处理器、256GB R-ECC DDR4 内存、2 个 Intel® I350

千兆网卡；软件环境为：64 位的 Ubuntu 14.04 LTS、Python 2.7.6、MySQLdb 1.2.5、libzmq 4.2.5、pyzmq 17.1.2、MySQL 5.5.61。

由于 NGAS 能够处理的标准 FITS 文件必须包含自定义的关键字 ARCFILE（ARCFILE = 'NCU.2003-11-11T11:11:11.111'），实验数据为 MUSER-I 的 40 万个已添加 ARCFILE 关键字的 FITS 文件，数据量约为 75.102GB（400000*201600B），存放在分配了 250GB 内存的 tmpfs（临时文件系统）中。

3.2 实验结果

基于 ZeroMQ 改进的多点观测数据存储与同步方法与 NGAS 中的数据存储与同步方法的实验结果性能对比如图 5 所示。订阅者使用基于 ZeroMQ 改进的数据存储与同步方法将这 40 万个 FITS 文件完全同步传输与存储下来所耗费的时间约为 333.834 秒（约 5.6 分钟），然而使用 NGAS 中的数据存储与同步方法所耗费的时间约为 13330.998 秒（约 222.2 分钟）。基于 ZeroMQ 实现的数据存储与同步方法比 NGAS 中的数据存储与同步方法快了约 39.993 倍。

3.3 讨论

实验结果表明基于 ZeroMQ 改进的 NGAS 的数据存储与同步方法在数据存储和同步方面性能明显优于 NGAS 的数据存储与同步方法，但是该方法也存在一些不足：

（1）由于基于 ZeroMQ 实现的 NGAS 数据发布端服务器要为每一个数据订阅者分配一个固定的端口和每个 IP 地址端口数为 65536 的限制，这就造成其只能为有限数量的订阅者提供服务；

（2）基于 ZeroMQ 的 NGAS 多点观测数据存储与同步方法的系统存在订阅端服务器因为无法及时存储高速接收的订阅数据而导致内存不足，进而导致订阅端服务器被杀掉。**在未来的工作中，我们将加入动态调整数据发布的机制来优化数据的同步传输效率。**

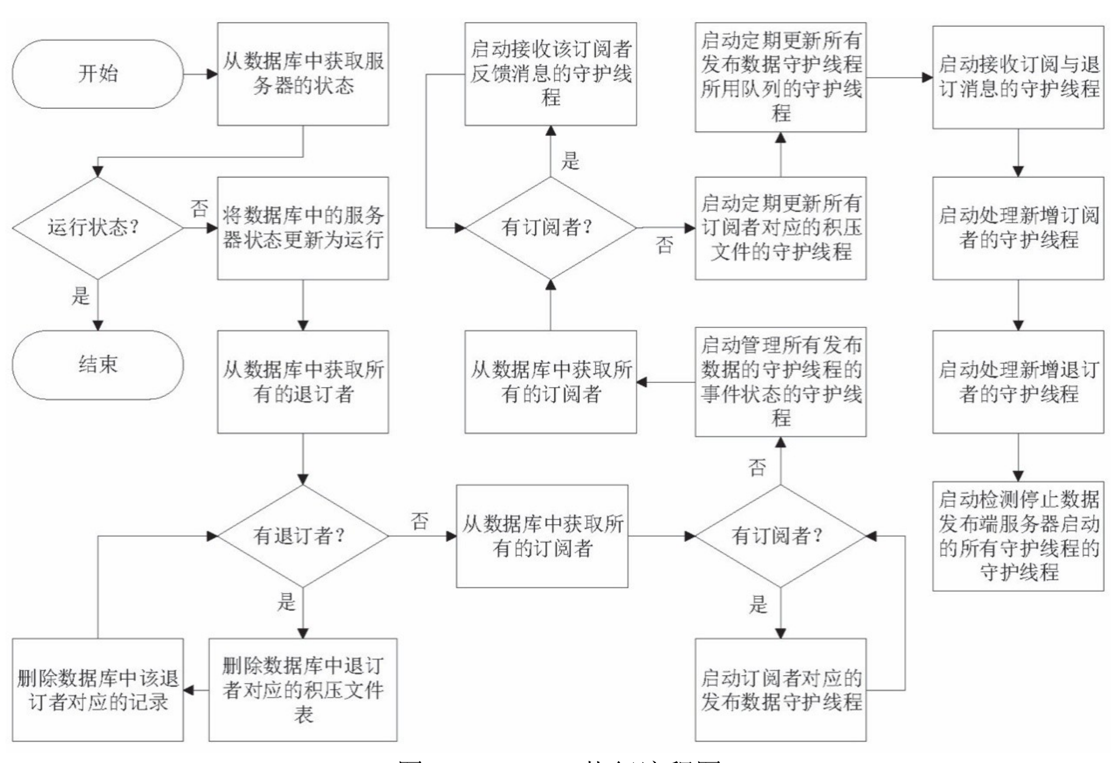


图 4 Pub-Server 执行流程图
Fig.4 A flowchart of Pub-Server

4 结束语

本文详细介绍了 NGAS 中的数据同步功能，并详细讨论了基于 ZeroMQ 改进的 NGAS 多点观测数据存储与同步方法及基于该方法实现的系统，通过实验验证了基于 ZeroMQ 改进的 NGAS 多点观测数据存储与同步方法实现的系统在数据**同步传输和存储效率方面性能明显优于 NGAS 中的数据存储与同步。接下来的工作中我们将会在更加真实的实验环境中测试新方法的远程数据同步性能和进一步优化其性能。本文的工作对 SKA 区域数据中心与 SKA 天文台数据中心之间的数据同步传输和存储有较好的参考价值。**

致谢：本文受到国家重点研发计划（2018YFA0404603，2016YFE0100300），国家自然科学基金委员会-中国科学院天文联合基金资助重点项目（U1831204），国家自然科学基金委员会-中国科学院天文联合基

金资助项目(No.U1831204, U1531132, U1631129)，国家自然科学基金资助项目(No.11403009, 11463003, 11773012)，广州大学“创新强校工程”项目（2017KZDXM062），云南省应用基础研究项目(2017FB001)，赛尔网络下一代互联网技术创新项目（No.NGII20170204）的资助。感谢国家天文台-阿里云天文大数据联合研究中心对本项工作的支持。

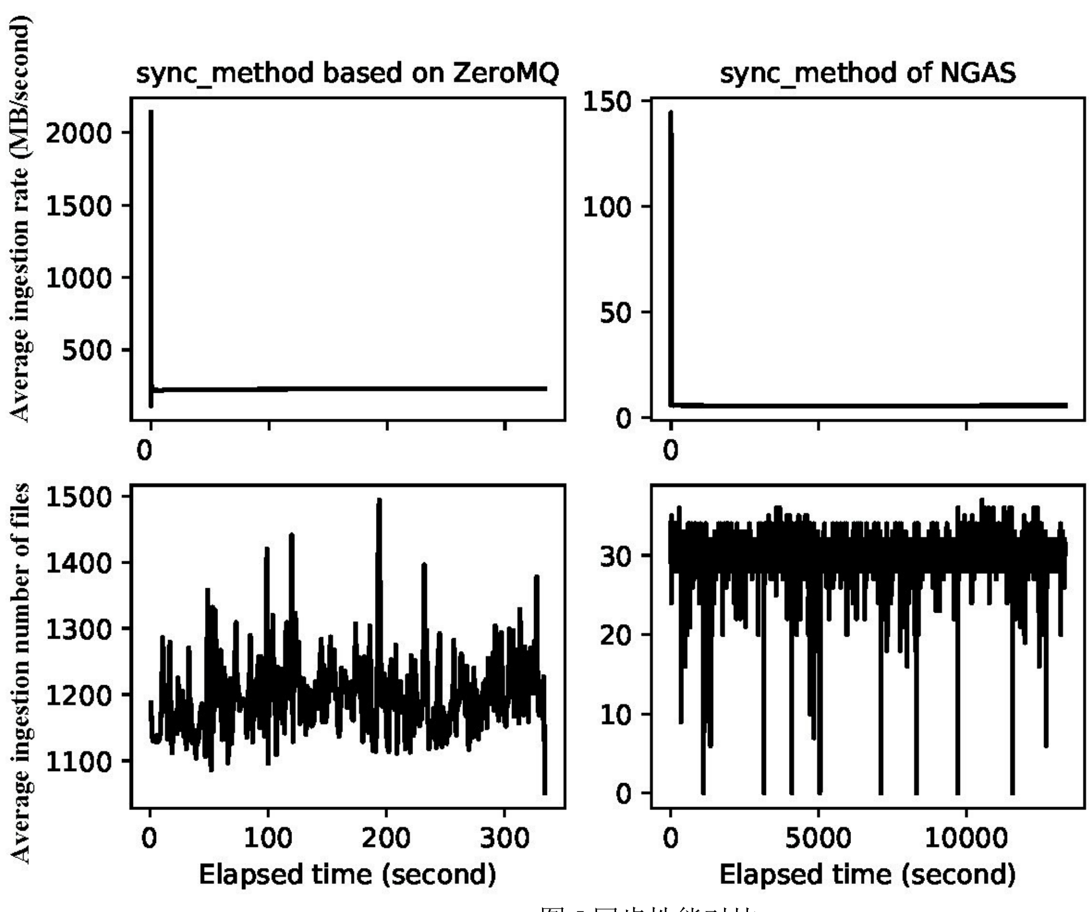


图 5 同步性能对比
Fig.5Synchronization performance comparison

参考文献

[1] 彭勃, 金乘进 杜彪, et al. 持续参与世界最大综合孔径望远镜 SKA 国际合作 [J]. 中国科学: 物理学, 力学, 天文学, 2013, 42(12): 1292-1307.
[2] Braun R, Keane E, Bourke T, et al. Advancing Astrophysics with the Square Kilometre Array [J]. PoS, 2015, 174.
[3] Dewdney P E, Hall P J, Schilizzi R T, et al. The square kilometre array [J]. Proceedings of the IEEE, 2009, 97(8): 1482-1496.
[4] 严俊. 天文与天体物理研究现状及未来发展的战略思考 [J]. 中國科學院院刊, 2011, 26(5): 487-495.
[5] Hall P, Schillizzi R, Dewdney P, et al. The square kilometer array (SKA) radio telescope: Progress and technical directions [J]. International Union of Radio Science URSI, 2008,236: 4-19.
[6] Broekema P C, Van Nieuwpoort R V, Bal H E. The Square Kilometre Array science data processor. Preliminary compute platform design [J]. Journal of Instrumentation, 2015, 10(07): C07004.
[7] Wicenec A, Knudstrup J, Johnston S. ESO's Next Generation Archive System [J]. The Messenger, 2001, 106: 11-13.
[8] Wu C, Wicenec A, Pallot D, et al. Optimising NGAS for the MWA Archive [J]. Experimental Astronomy, 2013, 36(3): 679-694.
[9] Harrison P, Knudstrup J, Wicenec A, et al. Implementing a VOStore Interface for NGAS; proceedings of the Astronomical Data Analysis Software and Systems XV, F, 2006 [C].
[10] Wicenec A, Knudstrup J. ESO's next generation archive system in full operation [J]. The Messenger, 2007, 129: 27-31.
[11] Wicenec A, Chen A, Checcucci A, et al. The ALMA Front-end Archive Setup and Performance; proceedings of the Astronomical Data Analysis Software and Systems XIX, F, 2010 [C].
[12] Stoehr F, Lacy M, Leon S, et al. The ALMA archive and its place in the astronomy of the future; proceedings of the Observatory Operations: Strategies, Processes, and Systems V, F, 2014 [C]. International Society for Optics and Photonics.
[13] Hintjens P. ZeroMQ: messaging for many applications [M]. " O'Reilly Media, Inc.", 2013.
[14] Hintjens P. Ømq-the guide [J]. Online: http://zguide/ zeromq org/page: all, Accessed on, 2011, 23.

NGAS-Based Multi-site Observation Data Storage and Synchronization Method

Shi Congming[1,3],Deng Hui[2,3], Dai Wei[3,4],Wei Shoulin[3],Wang Feng[1,2,3,4]

（1. Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650093, China; 2. Center for Astrophysics/Institute of Physics and Electronic Engineering, Guangzhou University, Guangzhou 510006, China; 3. Key Laboratory of Applications of Computer Technology of the Yunnan Province, Kunming University of Science and Technology, Kunming 650504, China;4. Yunnan Observatories, Chinese Academy of Sciences, Kunming 650216, China）

Abstract: The SquareKilometre Array (SKA) will have ultra-high sensitivity, ultra-fast survey speed and wide field of view, resulting in super-massive raw observation data.Massive data synchronization/delivery between the SKA observatories and regional data centers in various countries has been a difficult problem in the current SKA construction. The Next Generation Archive System (NGAS) used by SKA precursorexists inefficiency and insufficient performance in the application measurement. In this paper, we proposed aZeroMQ-based data storage and synchronization method. By adopting more efficient asynchronous messaging mechanism to realize synchronous data delivery, it can avoid the limitations of the HTTP protocol adopted by the NGAS. Experimental results showed that the new method is nearly 40 times faster than the original method used by NGAS in terms of average data archival/storage efficiency,can basically meet the requirements of the full-speed transmission of 10 GB bandwidth and has achieved a better use effect.

Keywords:NGAS; SKA; Storage and synchronization; Massive data; ZeroMQ